# **选题一**

### **肝脏基因表达谱芯片分析(人)**

### 背景介绍

HMG-CoA还原酶抑制剂(他汀类药物)的使用与潜在糖尿病危险因素(如肥胖、高血压和血脂异常)患者的糖尿病发病之间有一定的关系。

### 数据来源

GEO数据库： <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE130991>

部分已经整理好的数据链接：<https://jbox.sjtu.edu.cn/l/pn3myf>

### 数据介绍

本数据来自910名表现出不同程度肥胖和葡萄糖耐受性的受试者的肝脏组织样本的全基因组转录组谱，研究了他汀类药物治疗与葡萄糖稳态的关系。

每个受试者采集了一些临床数据，这些数据如下图所示：

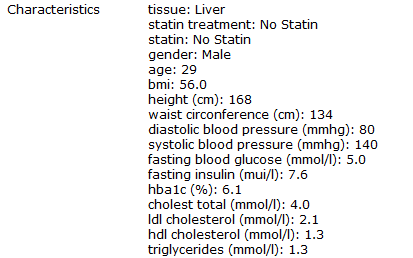


图1. 受试者临床数据示例

每个受试者对肝脏组织样本利用[HTA-2\_0] Affymetrix Human Transcriptome Array 2.0基因芯片检测了基因的表达信息，该芯片平台的详细信息见：[链接](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?mode=raw&is_datatable=true&acc=GPL20265&id=7661&db=GeoDb_blob128)。

对该数据的更多信息请阅读参考文献：<https://bmcmedgenomics.biomedcentral.com/articles/10.1186/s12920-019-0536-1>

### 探索研究以下问题：

1. 如果仅考虑临床因素，试用其他的临床变量来对血糖水平进行分类预测建模研究（分类标准为：fasting glucose<=5.5，5.6-6.9，>=7.0），并研究下空腹血糖水平与这些临床变量的关系（建模过程中的异常值、缺失值酌情处理）。
2. 比较三组之间的基因表达的差异，哪些基因的差异表达会潜在的造成空腹血糖水平的升高？你认为是什么原因？提出你的见解并从生物信息的角度来给与科学的解析。（差异表达基因给了参考的表格，你可以使用这个参考列表也可以自己做一遍差异表达分析的过程）
3. 挑选你认为有代表性的基因，建立分类预测模型，预测受试者的空腹血糖水平（高、中、低）。
4. 对问题3，可否建立回归模型预测受试者的空腹血糖值？
5. 将临床变量与你挑选的基因一起建立分类预测模型，该模型与前面的两个模型进行比较，提出科学问题以及解释。

其他参考信息:

1. 下载GEO表达谱数据的expression matrix文件中包含了所有样本的采集的临床信息信息，建议使用grep等命令提取grep '!' GSE130991\_series\_matrix.txt,对提取后的数据进行清洗、整理。
2. 直接下载的表达谱矩阵第一列为探针的名字，需要自己根据芯片平台的注释信息将其转换为对应的基因的名字，芯片平台的注释文件有一些house keeping的control探针，分析过程中请忽略。

# **选题二**

### **类风湿关节炎(RA)相关手部关节损害的程度评估**

### 背景介绍

类风湿性关节炎会累及关节，目前的量化办法是从放射图中的手、手腕和脚来计算，但是这种计算方法比较耗时、昂贵且有主观性。

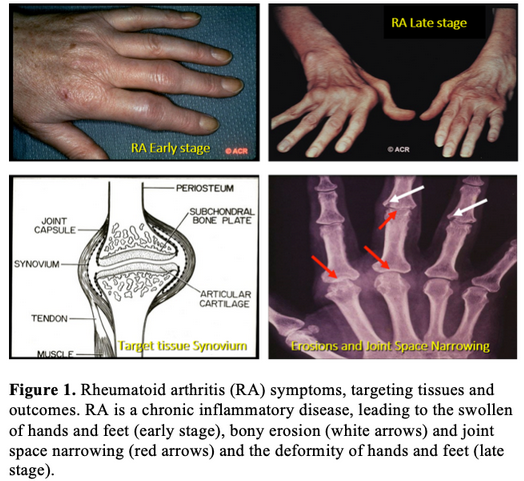


图2 风湿性关节炎患者手部关节受累示意（fusion and narrow）

### 数据来源

CLEAR知识库和TETRAD研究。数据链接：[E2-images.tar.gz   
  https://jbox.sjtu.edu.cn/l/H1NwMU](https://jbox.sjtu.edu.cn/l/0HRRQk" \o "E2-images.tar.gz)

### 数据介绍

本数据来自368名表现出不同程度手部关节损伤的类风湿性关节炎患者的放射图像，共728张图，有的图来自患者左右两只手，有的仅有一只，为简化项目，将两只手视为两个样本，本次研究中已经将每个关节受累程度的打分值进行了加和（label.txt的第二列为加和值）。

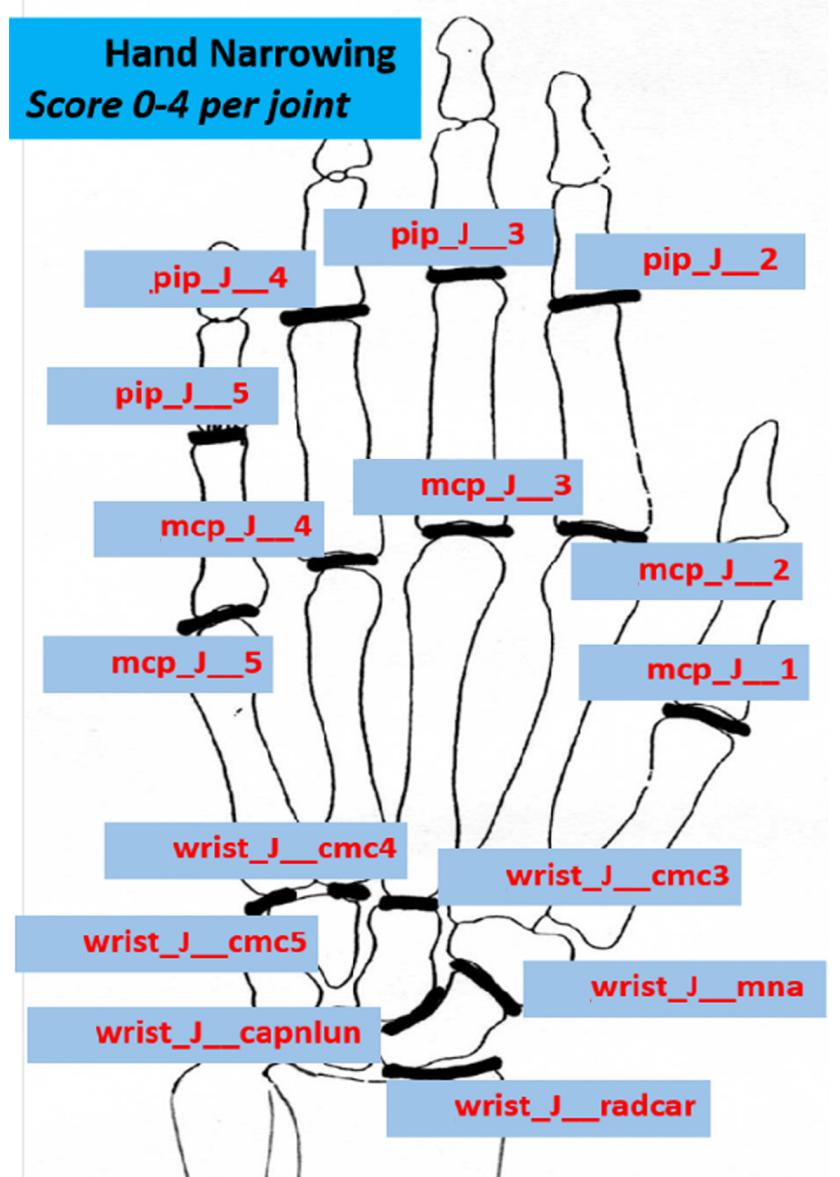


图3 本数据采集的手指关节部位指示图

### 探索研究以下问题：

1. 试建立分类模型，探讨利用深度卷积神经网络对手关节受累与否的预测效果
   1. 用所有关节受累打分值加和作为评估手关节是否受累的指标
2. 根据1获得的模型讨论下该模型是否合适？请探讨研究是否可以用深度卷积神经网络预测某个关节是否受累？
3. 请尝试对该数据集做拓展研究

# **选题三**

### **高通量显微镜图像中的蛋白质亚细胞定位评估**

### 背景介绍

绿色荧光蛋白GFP是实验室常用的蛋白标记物，通过显微镜观察荧光染色物的位置可以获知被标记蛋白的亚细胞定位。更多信息[..](https://wenku.baidu.com/view/b67b896202768e9951e73882.html)

### 数据来源

数据下载链接：<https://jbox.sjtu.edu.cn/l/SHwJtx>

每张显微图中仅有一个细胞，图中荧光蛋白的亚细胞定位标签从0~11分别代表：

|  |  |
| --- | --- |
| **label** | **label\_idx** |
| cell periphery | 0 |
| cytoplasm | 1 |
| endosome | 2 |
| er | 3 |
| golgi | 4 |
| mitochondrion | 5 |
| nuclear periphery | 6 |
| nucleolus | 7 |
| nucleus | 8 |
| peroxisome | 9 |
| spindle pole | 10 |
| vacuole | 11 |

### 数据介绍

该数据集有两个通道(RGB颜色模式下B通道为0)，红色荧光蛋白(mCherry)与胞质定位从而标记细胞轮廓，绿色荧光蛋白(GFP)在3’端标记内源性基因，表征蛋白的丰度以及蛋白的亚细胞定位。整个数据集包含65000个训练图像，12500个验证图像，以及12500个测试集。

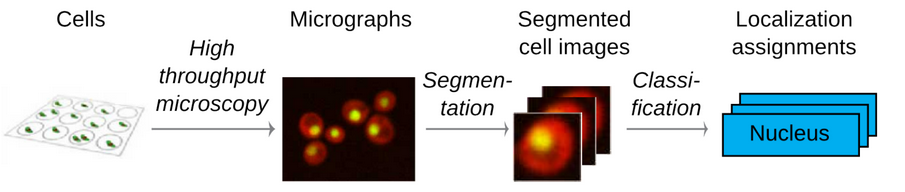


图4 来自文献的亚细胞定位研究方法

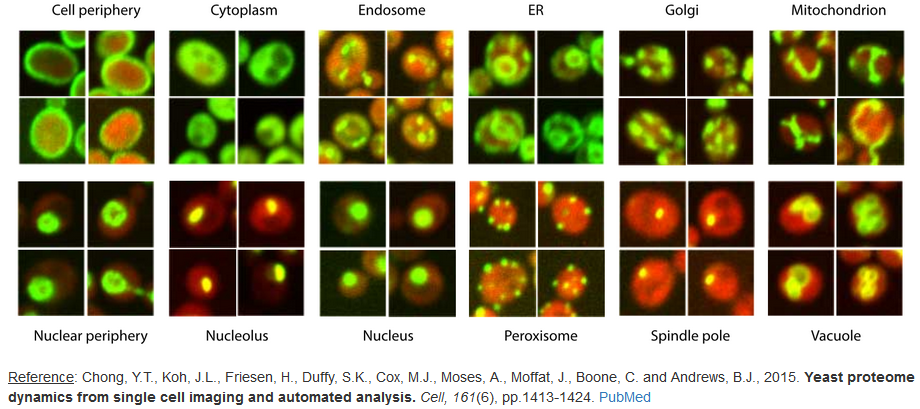


图5 不同类型的蛋白的亚细胞定位示例

### 探索研究以下问题：

1. 建立分类模型，预测GFP的亚细胞定位
2. 考虑一下纳入模型的训练集如果只用一个通道的数据与用2个通道的数据有什么区别?
3. 思考一下，你可以用红色荧光覆盖的区域来估计细胞的大小吗？（可选）

# **选题四**

### **糖尿病眼底损伤程度评估**

### 背景介绍

在临床医学上，眼科医生依据眼底疾病患者的眼底彩照图像进行详细的筛查与诊断并给出具体的治疗方案。通过对眼底的图像审查，医生可以或者病人是否得了青光眼（视杯视盘比例变化）、白内障、糖尿病眼底损伤等。

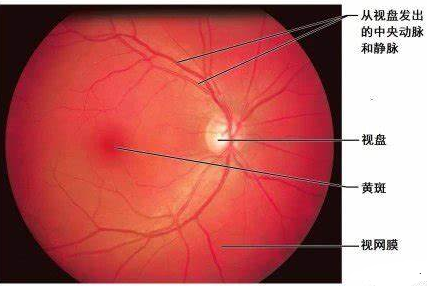


图6 典型的眼底图像

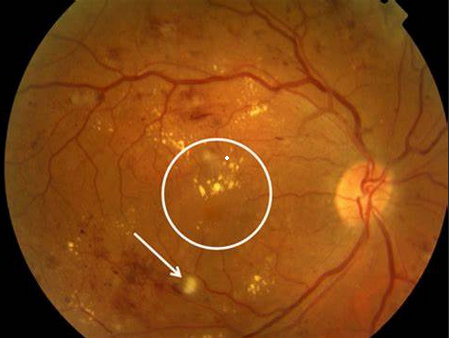


图7 糖尿病导致的眼底病变示例

### 数据来源

数据下载链接：<https://jbox.sjtu.edu.cn/l/B1xGxv>

数据包含了原始的图像以及划分好的training和validation set，每张图来自糖尿病患者的眼底图像，csv文件中给出了每张图的评价标准。请根据所提供的信息清洗数据，得到可用于训练的图像集及标签。

### 探索研究以下问题：

使用双眼视网膜眼底图像根据糖尿病性视网膜病变的严重程度对眼底图像进行分类建模（为简化研究不考虑是否来自同一人，左右两只眼睛视为不同样本即可）

思考你的模型的精准度可以达到医疗使用的程度吗？请阐述你的观点。

# **选题五**

### **基于序列突变信息的新冠病毒毒株聚类研究**

### 背景介绍

2019年出现的新型冠状病毒给全球经济、卫生、政治造成了严重的影响。一大批的科研机构贡献了大量的病毒全基因组检测样本，这些样本与参考病毒基因组进行比对后可以得到突变位点信息，基于突变的信息可以获知新冠病毒的变异株地域性聚集信息。

### 数据来源

数据下载链接：<https://jbox.sjtu.edu.cn/l/aoMe85> （内部数据，请勿外传）

### 数据介绍

本数据集包含346条高质量的新冠病毒测序的全基因组序列以及对应的突变检测数据(参考基因组：MN908947)。

### 探索研究以下问题：

1. 清洗整理基因组序列以及突变检测数据
2. 构造可用于科学研究的突变检测结果矩阵，用多种机器学习算法对突变信息进行聚类研究
3. 目前S蛋白是新冠病毒研究的热门蛋白，请就该蛋白对应基因上的突变位点单独进行研究，重复step2.

# **选题六**

### **急性胰腺癌患者ICU住院数据分析**

### 背景介绍

### 急性胰腺炎是一种相当严重的疾病，急性出血坏死性胰腺炎尤为凶险，发病急剧，死亡率高。

在现代临床医学中，呼吸机作为一项能人工替代自主通气功能的有效手段，已普遍用于各种原因所致的呼吸衰竭、大手术期间的麻醉呼吸管理、呼吸支持治疗和急救复苏中，在现代医学领域内占有十分重要的位置。呼吸机是一种能够起到预防和治疗呼吸衰竭，减少并发症，挽救及延长病人生命的至关重要的医疗设备。然而呼吸机属于资源并在许多医院甚至三甲医院的急诊科都不是特别的丰富，如果能提前预测到患者进入ICU后是否需要使用呼吸机，则可以提前优化医院呼吸机的分配额度，减少病人的等待时长，使病人尽可能的在病程最关键的阶段得到最有效的治疗。

### 数据来源

本题数据来自MIMICiii数据库，MIMIC-III是由麻省理工学院计算生理学实验室开发的公开可用数据集，其中包含了>60000次住院相关的数据，时间跨度为2001-2012年，包括人口统计学，生命体征，实验室测试、药物等。Nature上的一篇文献描述了该数据库（http://www.nature.com/articles/sdata201635）。

数据下载链接：<https://jbox.sjtu.edu.cn/l/G575Ug> （内部数据，请勿外传）

### 数据介绍

本题数据从MIMICiii数据库中提取筛选，筛选了一些变量。 变量释义如下：

|  |  |  |
| --- | --- | --- |
| **term** | **全称** | **释义** |
| alb | albumin | 白蛋白含量 |
| rdw | red blood cell distribution width | [红细胞分布宽度](http://mall.cnki.net/magazine/Article/ZZLC201904002.htm" \o "http://mall.cnki.net/magazine/Article/ZZLC201904002.htm) |
| pla | Platelets | 血小板 |
| tg | Triglyceride | [甘油三酯](https://www.baidu.com/s?wd=%E7%94%98%E6%B2%B9%E4%B8%89%E9%85%AF&tn=SE_PcZhidaonwhc_ngpagmjz&rsv_dl=gh_pc_zhidao" \o "https://www.baidu.com/s?wd=%E7%94%98%E6%B2%B9%E4%B8%89%E9%85%AF&tn=SE_PcZhidaonwhc_ngpagmjz&rsv_dl=gh_pc_zhidao) |
| NLR | Neutrophil lymphocyte ratio | [中性粒细胞淋巴细胞比值](https://wenku.baidu.com/view/606268b9bdeb19e8b8f67c1cfad6195f312be8a5.html" \o "https://wenku.baidu.com/view/606268b9bdeb19e8b8f67c1cfad6195f312be8a5.html) |
| gender | gender | 性别 |
| hs | Hospital stay | 住院时长 |
| ventilation | ventilation | 是否使用呼吸机 |

数据有缺失，请酌情处理。

### 探索研究以下问题：

1. 利用提取到的变量可否预测病人在住院期间是否会用到呼吸机？尝试建模并解释
2. 利用已知的变量是否可以预测病人的未来？（死亡or出院）
3. 能否利用提取到的这些变量来预测病人在ICU的住院时长？请尝试建模并解释
4. 讨论以上你建立的模型的可靠性，提出你的见解。

# **选题七**

### **吸烟和肺癌的因果关系研究**

### 背景介绍

肺癌是我国目前发病率和死亡率最高的恶性肿瘤。并且肺癌的发病率逐年攀升。

### 数据来源

数据下载链接：<https://jbox.sjtu.edu.cn/l/WuCIIu>（内部数据，请勿外传）

### 数据介绍

本题数据从TCGA中获取了肺癌患者的临床数据以及部分非肺癌患者的临床数据，并且下载了部分基因表达谱的数据，数据有缺失（临床数据与基因表达谱样本有不一致的请取交集），请酌情处理。

### 探索研究以下问题：

观测性的数据可以描述变量之间的相关关系，虽然研究表明了吸烟和肺癌有相关性，那么仅基于观测数据可以推测出来吸烟与肺癌有关系吗？

请从因果推断的角度来阐述你的观点。

[请于5月6日之前完成作业并发送到助教邮箱（kongyan@sjtu.edu.cn）。有疑问可以在微信群询问。](mailto:请于4月20日之前完成作业并发送到助教邮箱（kongyan@sjtu.edu.cn）。有疑问可以在微信群询问。)